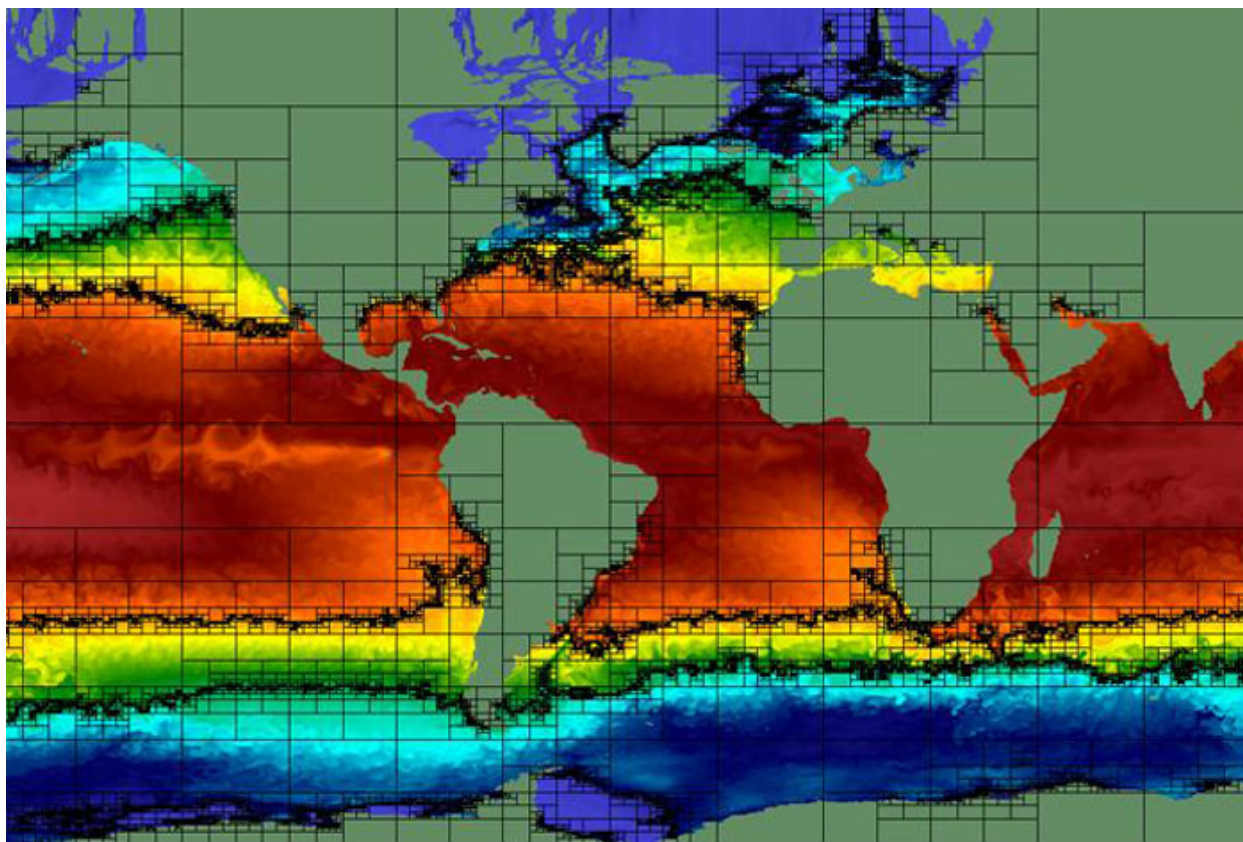# Data triage enables extreme-scale computing

August 1, 2014

The growing scale, size, and complexity of computing require prioritization to manage the data. However, resources are lacking to process all of the data fully, either by a computer or the end-user. To deal with the lack of resources, more important data need to be processed or examined first, and less important data are culled or ignored. Data selection and triage are important techniques for large-scale data, which can drastically reduce the amount of data written to disk or transmitted over a network. This is done through a data prioritization tree for arbitrary data sets. Therefore, Los Alamos scientists developed Analysis-Driven Refinement (ADR) visualization, a generalized analysis framework for ranking large-scale data. The researchers presented a whitepaper on the subject for the Big Data Exascale Computing workshop in Japan.

Significance of the research

The main focus for ADR is to prioritize data primarily generated by large-scale scientific simulations run on supercomputers. The Laboratory team designed ADR to be run in situ, so that data may be prioritized for triage operations that occur while the data is still in supercomputer memory. This strategy allows scientific data to be hierarchically partitioned in space, time and variable using user-defined, analysis-driven, importance measurements for data prioritization. Data triage algorithms at the extreme-scale determine what data are important for scientific analyses and decision-making.

As an additional benefit, the Analysis-Driven Refinement visualization enables the placement of automatic cameras. Typically the scientist must tell the visualization software where the data of interest are, and manually set the "virtual camera" of where the image is taken for the data. In ADR visualization, the cameras are automatically generated by detecting data of interest, without manual input by the scientist. This saves analysis time and could guide the user to important information.

Research achievements

An organization step occurs first, when partitioning ranks input data. A large-scale data set is recursively refined, or partitioned, into groups of data that have passed importance criteria. The relative importance data partitions are determined by the depth that the partitions occur in the data prioritization tree. Next, a selection step uses the resulting prioritization information to generate data products, such as sparse data sets or focus+context visualizations. In focus+context visualization, important data are highlighted for the user to save time or reveal data that may be obscured.

The researchers applied Analysis-Driven Refinement visualization to a POP (Parallel Ocean Program) simulation created by the LANL Climate, Ocean and Sea Ice Modeling team, which is part of LANL's and DOE's long-term climate studies. POP is a tenth of a degree, high-resolution eddy resolving simulation. The algorithm provided spatial partitioning using different values as importance for the ocean simulation data set. Moreover, the visualization identified virtual camera placements to highlight areas of interest to the scientists.

The research team

LANL researchers include Jonathan Woodring, Boonthanome Nouanesengsy, John Patchett, and James Ahrens of Applied Computer Science and Kary Myers of Statistical Sciences. The DOE Office of Science Advanced Scientific Computing Research (ASCR) and NNSA Advanced Simulation and Computing (ASC) funded the research, which supports the Laboratory's mission areas. In particular, the work benefits the Data Science at Scale and Computational Co-Design aspects of the Information Science and Technology science pillar.

Caption for image below: Camera placements produced using ADR visualization.